From the Department of Physiology and Pharmacology,

Karolinska Institutet, Stockholm, Sweden

FACILITATING PRECISION MEDICINE THROUGH ANALYSIS OF NEXT-GENERATION SEQUENCING PROJECTS

Qingyang Xiao

Karolinska Institutet

Stockholm 2020

Published by Karolinska Institutet.

Printed by Universitetsservice US-AB, 2020

ISBN 978-91-8016-030-8

Facilitating precision medicine through analysis of next-generation sequencing projects

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Qingyang Xiao

The thesis will be defended in public at Room Nils Ringertz, Floor 3 in Biomedicum, Karolinska Institutet, Solna, 2020/12/11 at 9.30 am

Principal Supervisor:
Assoc. Prof. Volker Lauschke
Karolinska Institutet
Department of Physiology and Pharmacology

Co-supervisor(s):
Assist. Prof. Isabel Barragan
Karolinska Institutet
Department of Physiology and Pharmacology

Opponent:
Prof. Ingolf Cascorbi
University of Kiel
Department of Experimental and Clinical Pharmacology

Examination Board:
Prof. Erik Eliasson
Karolinska Institutet
Department of Laboratory Medicine

Prof. Jan Dumanski
Uppsala University
Department of Immunology, Genetics and Pathology

Assoc. Prof. Nick Tobin
Karolinska Institutet
Department of Oncology-Pathology

“The first step is to establish that something is possible; then probability will occur.”

Elon Musk

ABSTRACT

Precision medicine constitutes an emerging strategy that aims at the individualization of healthcare by considering the personal molecular features and environmental factors of the patient in question. Genetic biomarkers constitute one dimension of a patient's molecular phenotype that can allow for treatment stratification. As such, incorporating genetic variability into clinical decision making has raised great interest with drug developers, regulators and in the wider medical community. Importantly however, most studies that evaluated associations of genetic variability with drug reponse or toxicity interrogated only selected, mostly common candidate variants and the prevalence and relevance of rare variants for pharmacogenetics remained largely unexplored. This thesis demonstrates how population-scale Next-Generation Sequencing (NGS) data can be leveraged to map the interindividual and ethnogeographic variability of genes with medical importance.

Papers I and II focused on ATP-binding casette (ABC) transporters, as an example of a pharmacogenetically relevant gene family, and show how their variability can have potential predictive value in breast cancer chemotherpy. The human ABC transporter family consists of 48 functionally important membrane proteins which mediate the active transport of a plethora of substrates, including a multitude of endogenous substrates as well as drugs, such as calcium channel blockers and various chemotherapeutics. Because of this physiological and clinical importance, **Paper I** systematically investigated the interindividual and ethnogeographic variability in the ABC transporter superfamily using NGS data of 138,632 unrelated individuals worldwide, and used an list of sophisticated computational algorithms to estimate their functional relevance. In total, 62,793 exonic variants were discovered, of which 98.5% were rare with minor allele frequencies (MAF) <1.5%. Based on these data, individuals were found to harbor between 9.3 and 13.9 deleterious *ABC* variants, only 0.3% of which were shared among all populations. As such, this work analyzed the landscape of *ABC* transporter variability on an unprecedented scale and revealed large interindividual and ethnogeographic variability with potential relevance for the treatment with ABC transporter substrates.

Paper II built on these findings by evaluating whether *ABC* transporter variability was associated with drug response. As drug resistance due to facilitated ABC transporter-mediated efflux of chemotherapeutics constitutes an important cause of morbidity and mortality, *ABC* transporter variability was evaluated whether it could predict treatment outcomes in breast invasive carcinoma (BRCA), clear cell renal carcinoma (ccRCC) and hepatocellular carcinoma (HCC). In contrast to previous studies, these analyses did not only consider common *ABC* polymorphisms but considered also rare genetic variants using mutational burden testing. Importantly, variant burden of *ABCC1* was found to significantly assoiate with reduced survival in BRCA patients, specifically in those subgroups treated with the MRP1 (the transporter encoded by *ABCC1*)

substrates doxorubicin (p=0.0088) and cyclophosphamide (p=0.0011). In contrast, no association was discovered in tamoxifen-treated patients (p=0.13). Multiple variants enriched in the high mutational burden group affected residues in functionally important transporter domains providing additional mechanistic support. Combined, these results argue for a model in which multiple variants with individually small effect sizes shape drug resistance, thus incentivizing a shift in strategy away from the interrogation of candidate variants and towards the incorporation of germline data for precision cancer medicine.

Paper III indicated how publically available sequencing data from individuals can be used to provide accurate estimates of population-specific carrier rates and genetic complexity of 450 human autosomal recessive (AR) diseases. Specifically, population-scale NGS data of individuals free from clinically diagnosed congenital disorders was used to identify disease allele carrier frequencies for 450 AR disorders. Using 85 diseases with known epidemiology, the data showed that our prevalence estimates corresponded well to clinically reported incidences ($p<0.001$; $R=0.68$). Furthermore, these data allowed for the first time to evaluate the genetic complexity of the human AR diseasome and estimate population-specific founder effects. As such, these analyses reveal the molecular genetics of AR diseases with unprecedented resolution and provide important insights into epidemiology, complexity and population-specific founder effects, which can provide a powerful resource for clinical geneticists to inform population-adjusted genetic screening programs, particularly in otherwise understudied ethnogeographic groups.

In conclusion, by utilizing sophisticated computational methods for the analysis of publically available population-scale sequencing data of >130,000 individuals, this thesis uncovered the landscape of genetic variability in genes with importance for pharmacogenetics and congenital disease. The resulting findings aspire to improve pharmacogenetic interpretations and carrier screening programs and, hopefully, can contribute to the advancement of precision medicine.

LIST OF SCIENTIFIC PAPERS

Papers/manuscripts **included** in the thesis frame:

I. **Qingyang Xiao**, Yitian Zhou, Volker M. Lauschke. **Ethnogeographic and inter-individual variability of human ABC transporters.**
Human Genetics, 2020, 139, 623–646.

II. **Qingyang Xiao**, Yitian Zhou, Stefan Winter, Florian Büttner, Elke Schaeffeler, Matthias Schwab and Volker M. Lauschke. **Germline variant burden in multidrug resistance transporters is a therapy-specific predictor of survival in breast cancer patients.**
International Journal of Cancer, 2020, 146, 2475-2487.

III. **Qingyang Xiao**, Volker M. Lauschke. **The prevalence, genetic complexity and population-specific founder effects of human autosomal recessive disorders.**
Genome Medicine [*Under review*]

Papers/manuscripts **not included** in the thesis frame:

I. Michäel Duruisseaux et al**. Epigenetic prediction of response to anti-PD-1 treatment in non-small-cell lung cancer: a multicentre, retrospective analysis.**
The Lancet Respiratory Medicine, 2018 Oct;6(10):771-781.

II. **Qingyang Xiao**, André Nobre, Pilar Piñeiro, Miguel-Ángel Berciano-Guerrero, Emilio Alba, Manuel Cobo, Volker M Lauschke, Isabel Barragán. **Genetic and Epigenetic Biomarkers of Immune Checkpoint Blockade Response.**
Journal of Clinical Medicine, 2020 Jan 20;9(1):286.

III. Juan Luis Onieva, **Qingyang Xiao**, Aurora Laborda-Illanes, Pilar Piñeiro, Alicia Garrido-Aranda, Elena Gallego, Cynthia Robles-Podadera, Rosario Chica-Parrado, Daniel Prieto, Vanessa De Luque, María José Lozano, Martina Álvarez, Pedro Jiménez, Alfonso Sánchez, Emilio Alba, Miguel Berciano-Guerrero, Javier Oliver, Manuel Cobo, Isabel Barragán. **B lymphocyte signature in melanoma patients predicts prognosis and response to Immune Checkpoint Blockade.**
Manuscript

IV. **Qingyang Xiao**, Yitian Zhou, Volker M. Lauschke. **The impact of variants in ATP-binding cassette (ABC) transporters on breast cancer treatment**. [Review]
Pharmacogenomics [*Accepted*]

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ABC	ATP-binding cassette
A1AT	Alpha1 anti-trypsine
ACB	African Caribbeans in Barbados
ACMG	American College of Medical Geneticists
ACOG	American College of Obstetricians and Gynecologists
AFP	Alpha-fetoprotein
AR	Autosomal recessive
ASW	Americans of African Ancestry in SW USA
BCRP	Breast cancer resistance protein
BEB	Bengali from Bangladesh
BRCA	Breast invasive carcinoma
ccRCC	Clear cell renal cell carcinoma
CDX	Chinese Dai in Xishuangbanna, China
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry
CF	Cystic fibrosis
CHB	Han Chinese in Beijing, China
CHS	Southern Han Chinese
CI	Confidence interval
CLM	Colombians from Medellin, Colombia
CPS1	Carbamoylphosphate synthetase I
DSS	Disease-specific survival
ESN	Esan in Nigeria
ExAC	Exome Aggregation Consortium
FDA	Food and Drug Administration
Fig	Figure
GBR	British in England and Scotland
GIH	Gujarati Indian from Houston, Texas
gnomAD	Genome Aggregation Database
GWD	Gambian in Western Divisions in the Gambia

HCC	Hepatocellular carcinoma
HGP	Human Genome Project
HR	Hazard ratio
IBS	Iberian Population in Spain
ITU	Indian Telugu from the UK
JPT	Japanese in Tokyo, Japan
KHV	Kinh in Ho Chi Minh City, Vietnam
LOF	Loss of function
LWK	Luhya in Webuye, Kenya
MAF	Minor allele frequency
MDR1	Multidrug resistance protein 1
MRP1	Multidrug resistance associated protein 1
MSL	Mende in Sierra Leone
MXL	Mexican Ancestry from Los Angeles USA
NBD	Nucleotide-binding domain
NGS	Next-Generation Sequencing
O/E	The number of observed variants / the number of expected variants
OMIM	Online Mendelian Inheritance in Man
PEL	Peruvians from Lima, Peru
PJL	Punjabi from Lahore, Pakistan
PUR	Puerto Ricans from Puerto Rico
SMRT	Single molecule real-time
SNP	Single nucleotide polymorphism
STU	Sri Lankan Tamil from the UK
TCGA	The Cancer Genome Atlas
TMD	Transmembrane domain
TSI	Toscani in Italia
WES	Whole exome sequencing
WGS	Whole genome sequencing
YRI	Yoruba in Ibadan, Nigeria

1 INTRODUCTION

Precision medicine describes a medical model that aims at the personalization of treatment, including the selection of therapeutic modalities and regimens, by taking genomic, proteomic and metabolomic signatures of the individual patient into account. Particularly genetic biomarkers have received great attention from the medical community in pursuit of precision medicine because they are inherently stable and relatively easy to probe. Particularly, precision genomic medicine has emerged in the fields of pharmacogenomics and genetic disease diagnostics [1].

However, commonly identified genetic candidate variants can only explain a small portion of the entirety of hereditary phenotypes. The remaining missing heritability constitutes a major problem in explaining genotype-phenotype association, and impedes the development of translational genetics for precision medicine. Therefore, this thesis evaluates the signature of rare genetic variability in pharmacogenetic and disease-associated genes and whether the consideration of such rare and often neglected variants using computational tools can improve prediction models and, eventually, facilitate precision medicine.

Specifically, the work is focused on two areas of genomic medicine, pharmacogenomics and genetic diseases:

Pharmacogenomics

It is estimated that 25%-50% of drug treatments do not result in the intended response or cause adverse events [2], 20-30% of which can be explained by genetic factors. Importantly however, twin studies suggest that commonly considered polymorphisms can only explain 30-40% of the heritable inter-individual differences in drug disposition [2]. For instance, up to 89% of caffeine pharmacokinetics were heritable after correcting for effects of smoking and hormonal contraceptives; however, only 8% of these differences were attributed to common variants in *CYP1A2* [3]. Similarly, the vast majority of metoprolol and torsemide pharmacokinetics variability is heritable, whereas known variants in genes implicated in their disposition explain only 2%-39% of this variance [4]. These results suggest that other genetic factors beyond the commonly interrogated variants might contribute to the observed differences in drug disposition.

Genetic diseases

Missing heritability is also relevant for our understanding of and screening for genetic disease. For instance, the commonly interrogated pathogenic variants in *ABCA4* can only explain 83.6% of the heritability of *ABCA4*-associated retinal dystrophy [5]. Similarly, the recommended carrier screening program by the American College of Obstetricians and Gynecologists (ACOG) and the American College of Medical Geneticists (ACMG) for cystic fibrosis (CF), comprising the 23 most common *CFTR* variants, only identifies around 80% of CF cases [6]. Combined, these data suggest that commonly considered well-characterized candidate single nucleotide polymorphisms (SNPs) can only explain a part of the heritable disease risk.

What explains the missing heritability? Evan Eichler and colleagues [7] suggested contributions of multiple factors:

1) Effects of gene-environment interactions rather than of genetic factors alone.

2) Epigenetics may contribute to the inheritance of phenotypes of interest.

3) Genetic variants interact with other biomolecules to affect protein expression.

4) Genetic variants of interest in unexplored genomic regions.

5) Effects of rare variants.

Apart from the missing heritability, another challenge in the development of precision genomic medicine is the population bias in current genomic research databases and references, which impairs the understanding of human diseases and aggravates healthcare imbalances in minority populations [8]. For instance, the GWAS Catalog, containing 35 million samples from 2,511 studies, is constituted predominantly of individuals of European descent (81%) and only 14% of Asian descent and 5% from rest populations worldwide [9]. Similar disparity of population composition is observed in the Allele Frequency Aggregator database (https://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/) and GTEx, of which 84% and 85% of the included individuals are of European descendent, respectively [10]. This disproportionate population makeup of the databases and references may impede analyses of the genetic disease aetiology and complicate treatment stratifications for most non-European populations. As only one example, *CYP2D6*10*, an allele impairs the enzyme activity and reportedly associated with toxicity of risperidone [11], is harboured by 70% of Asians yet <10% in worldwide other populations [12]. The vast difference of clinical important allele frequencies among populations indicate that population-tailored pharmacogenetic testing programs may be more efficient than if

they were based on global guidelines. Combined, the biased genomic sampling can impair the implementation of precision medicine in specific populations.

This PhD thesis aims to extend pharmacogenetic and genetic disease analyses beyond the analyses of commonly interrogated genetic variants by incorporating comprehensive genomic profiles based on population-scale Next-Generation Sequencing (NGS) data. In the following sections, I will provide information about how NGS technology has revolutionized human genomics, depict the trajectory of human NGS projects and illustrate their application for precision medicine.

1.1 SEQUENCING TECHNOLOGIES

1.1.1 First generation sequencing

First generation sequencing refers Sanger sequencing and Maxam-Gilbert sequencing [13]. Sanger method employs dideoxynucleotides to terminate chain and determine the sequence. Sanger sequencing is highly robust and was the standard genomic approach before the advent of NGS for three decades. Despite the accuracy, sanger sequencing generates low throughputs and is costly and time-consuming. Combined, sanger sequencing gradually becomes as a means of validation rather than an explorative method for biological question of interest in general.

1.1.2 Second generation sequencing

Since the early 21st century, sequencing technologies have developed rapidly. Compared to first generation sequencing, these new technologies feature higher throughput and sensitivity, as well as broader genomic coverage [14–17]. Second generation sequencing, often referred to as NGS technologies, were specifically developed for genetic, epigenetic and gene expression research across a wide range of biological applications. Among these, whole exome sequencing (WES) and whole genome sequencing (WGS) are two common utilized NGS techniques in genome research. WGS is capable of decoding the sequence of the entire genome while WES only focuses on variants in the protein-coding region, covering only approximately 1.5% of the whole genome [18].

Importantly, NGS mostly generate short reads (< 1 kb) [19]. Short-read data is overall accurate, can sequence loci with high depth and is supported by a multitude of bioinformatics tools [13,20], which has resulted in short-read NGS becoming the current standard tool in sequencing-based clinical diagnostics [21]. However, this

method has limited sequencing accuracy for repetitive sequences [22] or regions with high genomic complexity, including regions with high GC content, structural variations or highly homologous nearby genes [22]. These problems are of particular importance for pharmacogenomics, as a multitude of clinically relevant genes are genetically complex, including *CYP2D6*, *CYP2B6* and various *HLA* genes.

1.1.3 Third generation sequencing

To overcome these limitations, long-read sequencing, often referred to as third generation sequencing, has been developed, which is capable of generating reads of many kbs in length [23]. These technologies facilitate the interrogation of loci that are inaccessible by conventional short read sequencing [22]. The leading two platforms comprise single molecule real-time (SMRT) and nanopore sequencing [20]. Both methods increase the genetic coverage of sequencing and mapping accuracy in difficult sequences and allow single molecule haplotype profiling [24]. These technical advantages have facilitate the clinical diagnosis of a variety of diseases, such as Huntington's disease [25] and Fragile X syndrome [21]. Similarly, the technology demonstrates other potential clinical utility by precisely promotes the variant and haplotype interpretation of drug toxicity related genes such as *HLA* genes and *CYP2D6* [26]. Additionally, it's also anticipated that the technology will have a dramatic impact of variant interpretation of population sequencing projects [24]. Main disadvantages of third generation sequencing include high cost, high requirement for expensive materials and immature bioinformatics support [24].

1.2 EVOLUTION OF HUMAN SEQUENCING PROJECTS

1.2.1 Human Genome Project

The Human Genome Project (HGP) was initiated in 1990 and finished in 2003 with the aim to decode the entire human genome. Overall, the HGP covered >99% gene-containing regions of the human genome with >99% sequencing accuracy. The project completed with a human reference genome comprising 2.85 billion nucleotides and estimated approximately 20,000-25,000 protein-coding genes across the human genome [27]. This therefore lay a substantial foundation for biomedical and medical research.

1.2.2 The 1000 Genome Project

Following the reference genome, the 1000 Genome Project was initiated to investigate the genotypic diversity across worldwide populations [28]. The project overall included

2,504 individuals from 5 superpopulations subdivided into 26 well-defined subpopulations, and provided allele frequency and haplotype information for a total of 84.7 million SNPs and 36.7 million indels. The project demonstrated for the first time the drastic genomic heterogeneity across ethnogeographic groups (Fig. 1).

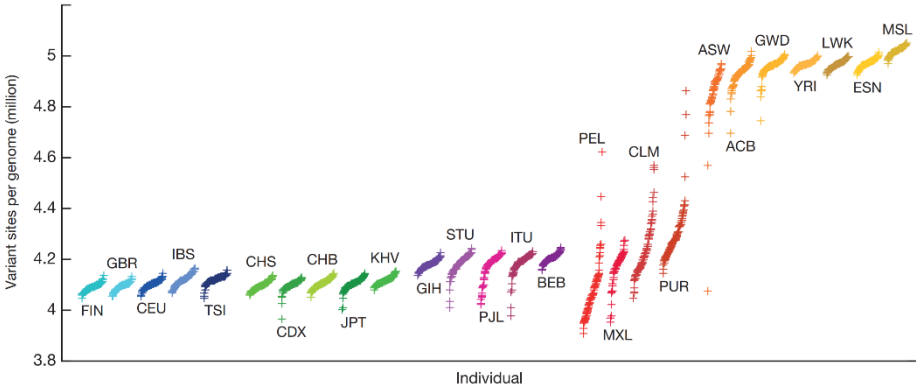

Fig 1. Number of variant sites of each subpopulation individual in 1000 Genome Project by subpopulation. *(Reprinted from* [29]*)*

1.2.3 Exome Aggregation Consortium (ExAC) and The Genome Aggregation Database (gnomAD)

While the 1000 Genomes Project analyzed well defined populations, the overall cohort size was relatively small and as such not suitable to adequately capture rare genetic variants. Therefore, the Exome Aggregation Consortium (ExAC) formed with the goal to unify available sequencing data from a variety of projects into one central database that eventually contained exome sequences of 60,706 individuals [30]. Due to the increased number of individuals, this dataset provided an improved resource for the analysis of rare genetic variability with relatively high resolution.

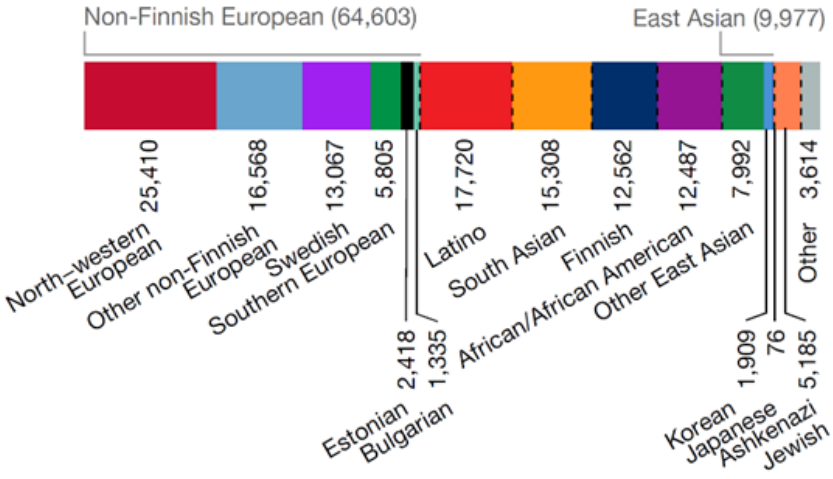

Fig 2. Population composition of gnomAD cohort. *(Reprinted from* [31]*)*

ExAC was further extended to include increasing numbers of WGS data in the form of The Genome Aggregation Database (gnomAD) with substantially enlarged cohort size to 141,456 individuals (Fig. 2), including 12,487 Africans, 17,720 Latinos, 5,185 Ashkenazi Jews, 9,977 East Asians, 64,603 non-Fin Europeans, 12,562 Finns and 15,308 South Asians [31]. Furthermore, the latest gnomAD version (v3) additionally included genomic information of the Amish, as an important founder population.

The project overall identified 14.9 and 229.9 million high-confidence variants for exome and genome datasets, respectively, as well as 433,371 structural variations [32]. Among these, 443,769 variants lead to frameshifts, stop gain and splicing site disruption across the human genome using stringent filtering criteria, suggesting the genome-wide existence of putative loss-of-function (LOF) variants [31]. Additionally, gnomAD estimated the LOF intolerance of each gene by calculating the ratio of the number of observed pLOF variants to the number of statistically expected pLOF variants (O/E). These genes with lower O/E value tend to be intolerant to pLOF variants and align with genes, the mutations of which are known to result in Mendelian disorders (Fig. 3). Taken together, gnomAD constitutes a reliable large-scale resource for population genetics.

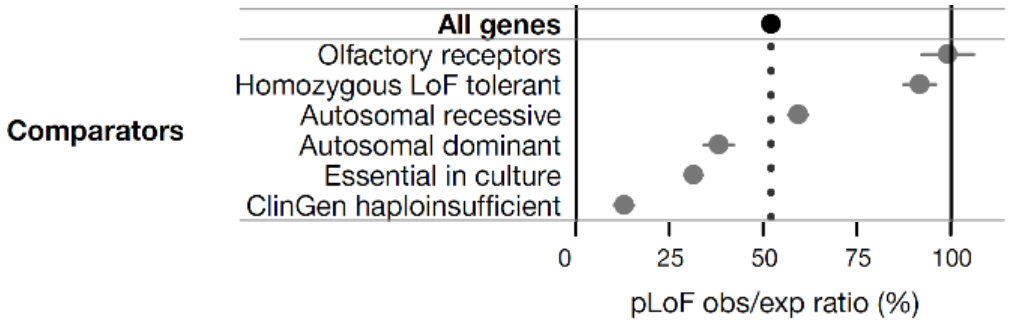

Fig 3. Constraint scores of each gene category with 95% confidence interval. *(Modified from* [33]*)*

1.3 BIOMEDICAL APPLICATIONS OF SEQUENCING DATA

1.3.1 Pharmacogenomics

1.3.1.1 ***Introduction into pharmacogenomics***

Up to 15% of EMA-approved and 264 FDA-approved medicines contain pharmacogenomic information in their labels, 7% of which are clinically actionable, affecting 18% of all drug prescriptions in the US [34–36] (Fig. 4). The main goal of pharmacogenomics is to identify how the genetic makeup of an individual impacts drug pharmacokinetics, response or toxicity of drugs. For instance, carriers of the *HLA-B*57:01* allele are at risk to develop abacavir hypersensitivity syndrome, whereas non-carriers are completely protected. As a consequence of this discovery, preemptive pharmacogenetic testing for *HLA-B*57:01* is compulsory before abacavir prescription. Similarly, rs4363657 in the *SLCO1B1* gene, encoding the OATP1B1

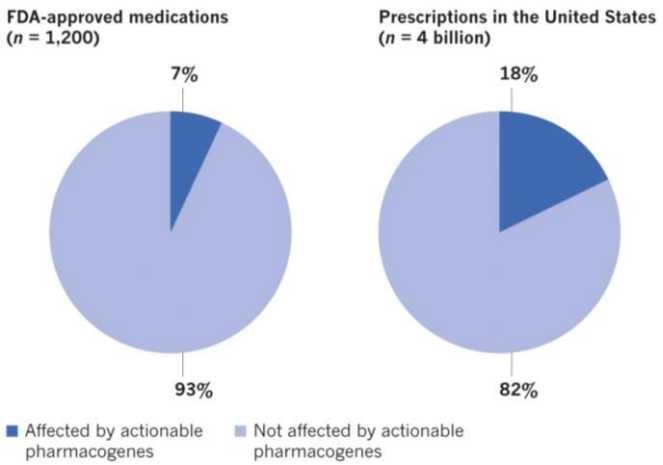

Fig 4. Drug/prescription classification according to actionable germline pharmacogenetics. *(Reprinted from* [36]*)*

simvastatin transporter, has been identified as a strong predictor of simvastatin-induced myopathy with odds ratio of 17.4 in variant homozygotes [37], and these findings have resulted in the abolishment of high dose prescriptions of simvastatin throughout Europe. Nowadays, more pharmacogenetic tests are available if the prescribing physician wants to evaluate the pharmacogenetic risk of potential toxicity preemptively. The testing is most commonly utilized in the pharmacological treatment of depression, anxiety, lipid disorders and hypertension, however, interestingly, the number of single gene test prescription in the US was declining over the last few years [38].

1.3.1.2 ***Pharmacogenetic testing in cancer chemotherapy***

One major arena of pharmacogenetic testing is the support of clinical decision-making in cancer chemotherapy. Notable examples are genotyping of *DPYD*, *UGT1A1* and *TPMT* for the guidance of fluoropyrimidine, irinotecan and mercaptopurine prescriptions. For *DPYD,* the most extensively studied variant **2A* (rs3918290, splice donor) as well as others [39–41], have been reported to significantly reduce dihydropyrimidine dehydrogenase activity, resulting in reduced fluoropyrimidine detoxification and increased risk of life-threatening toxicity. Based on these findings, the preemptive testing of functionally relevant *DPYD* variability is recommended when considering treating colon cancer patients with fluoropyrimidines in the Netherlands. Similarly, genotype-guided prescribing following preemptive testing of *TPMT* variants can reduce the risk of thioguanine toxicity [42]. In addition, *UGT1A1*28* homozygotes are poor metabolizers of irinotecan and at increased risk to develop severe toxicity after treatment [43], resulting in the US FDA recommending irinotecan dose reductions for *UGT1A1*28* carriers.

1.3.1.3 ***ABC polymorphism in cancer treatment***

While the aforementioned genetic variants constitute strong predictors of drug-specific toxicity, drug resistance constitutes arguably an even larger problem in clinical oncology. Despite the dramatic progress of the industry of next generation anticancer drug development, 90% of patients are primarily resistant or acquire resistance during cancer chemotherapy [44], which is the major cause to the treatment failure. Common resistance mechanisms include drug efflux, change of drug targets, limited drug activation or facilitated drug inactivation [45].

Excessive drug efflux causing chemotherapy resistance is primarily contributed by ATP-binding cassette (ABC) transporter superfamily (Fig. 5), pumping the drug substrates into intracellular or extracellular space in an ATP-dependent manner [46]. The human ABC superfamily consists of 48 functional *ABC* transporters attributed to 8 subfamilies and 22 pseudogenes [46,48], of which 12 have been found to mediate the transport of chemotherapeutic drugs [49], of which the MDR1-encoding *ABCB1*, MRP1-encoding *ABCC1*,and BCRP-encoding *ABCG2* are the most comprehensively studied ABC transporters in cancer chemotherapy to date. ABC transporters are

widely expressed in liver, placenta, kidney, blood-testis barrier and blood-brain barrier in mammals [46].

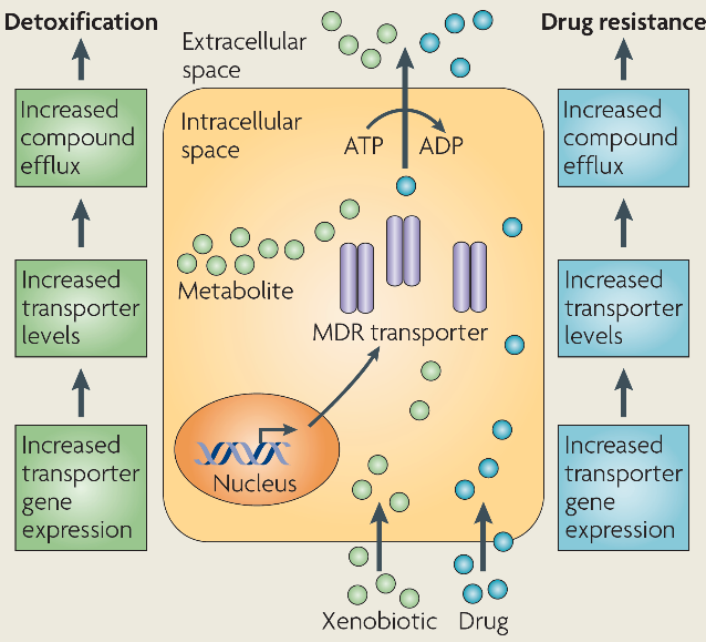

Fig 5. ABC transporter related multidrug resistance mechanism. *Increased expression of certain ABC transporters will increase the pump-out of compound from cytoplasm. The reduced drug concentration inside the cell will therefore decreased the toxicity and response simultaneously. (Reprinted from* [47]*)*

Genetic variants of *ABC* transporters reportedly can affect substrate transport activity *in vitro*. For instance, the synonymous variant I1145I of *ABCB1* affects cotranslational folding of MDR1 through the introduced rare codon, resulting in reduced levels of functional transporter [50]. Clinically, genetic variants of at least 12 *ABC* transporters reportedly significantly associated with drug response and toxicity in cancer medicine. *ABCB*, *ABCC* and *ABCG* subfamilies were most extensively investigated *ABC* subfamilies while *ABCA*s and others are scarcely found pharmacogenetic associations. Specifically, rs1128503, rs2032582 and rs1045642 of *ABCB1* are so far the most extensively investigated variants and discovered affect the drug response or toxicity for of chemotherapy for at least 7 cancers.

Importantly, rare variants with MAF < 1%, rather than above mentioned common variants, occupy the vast majority (92.9%) of genetic variations of pharmacogenes (Fig. 6) [51]. Nevertheless, rare variants conventionally draw little attention to medical community due to the low prevalence and cost-effectiveness of research funding. In addition, rare genetic variants reportedly contribute 30-40% of pharmacogenetic functionality variability. More detailed analyses of other pharmacogene families have revealed the abundance of rare variants and their functional importance in CYPs [52], SLC [53] and SLCO transporters [54]. These evidences together suggest that rare variants can be of importance for understanding the interpersonal variability of ABC transporter activity in the era of precision medicine. In contrast, most studies only focused on a limited set of *ABC* common variants and analyzed only relatively small cohorts with unclear population definition [55–58]. Such study designs decrease the statistical power and hinders the clinical interpretation of *ABC* transporter variability in worldwide diverse populations.

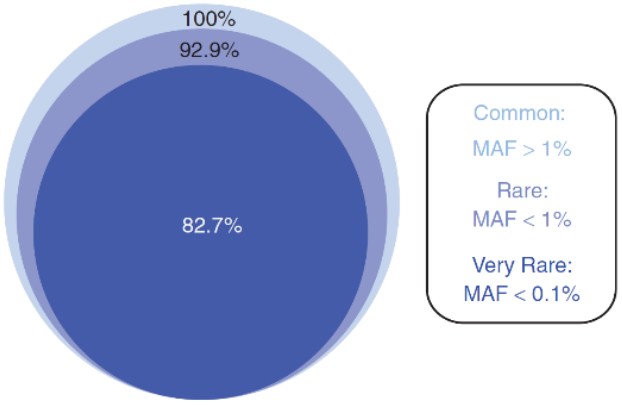

Fig 6. Variant composition of pharmacogenes by MAF. *(Reprinted from* [51]*)*

1.3.2 Pathogenic variants underlying genetic diseases

1.3.2.1 ***Introduction into congenital disease genetics***

It is estimated that individuals are up to 99.9% genetically identical, with only approximately one variation every 1,000 bases between individuals [59]. Among variant classes, SNPs are most abundant accounting for approximately 90% of human genetic variation [60]. Apart from SNPs, genetic variants include indels and structural variants. Indels are defined as small insertions or deletions with size < 1kb [61] and are of widespread molecular and phenotypic importance [62,63]. Furthermore, several indels constitute well-characterized causes of Mendelian disorders, including CF [64], Fragile X syndrome [65] and Huntington's disease. Likewise, a multitude of larger structural variants defined as >1kb [61] can underlie the molecular pathogenesis of genetic disease, including Hunter syndrome, Williams-Beuren syndrome and Sotos syndrome [66–68]. In general, the protein-coding region variants are generally considered as more clinically relevant than intronic variants.

Virtually all variants causing congenital diseases are rare, commonly defined as MAF<1%, whereas common variants have been mostly implicated in complex disease risk, for instance in type 2 diabetes or schizophrenia [69]. Notably however, such common variants have generally low effect size [70]. In contrast, rare genetic variants often have larger effect size, as shown for instance for type 1 diabetes and Alzheimer's disease [71,72], as well as for hypercholesterolaemia and nonsyndromic hearing loss [73,74].

1.3.2.2 ***Carrier screening of autosomal recessive disorders***

Autosomal recessive (AR) disorders are genetic diseases that result from homozygosity or compound heterozygosity of pathogenic variations. To date, more than 2,000 AR disorders have been identified worldwide [75] with 3 per 1,000 neonates being affected by AR disorders recessive disorders globally [76] and around approximately 35% individuals are carrier for at least one Mendelian disorder [77]. Notably, this incidence can be even substantially higher in communities that are genetically isolated, such as Ashkenazi Jews or Old Order Amish, or populations where consanguinity is common [78,79].

Carrier screening aims to identify if clinically asymptomatic persons are carriers of pathogenic variants [80]. This information can thus facilitate genetic reproductive counselling and thereby contribute substantially to public health.

1.3.2.2.1 Carrier screening for monogenic autosomal recessive disorders

The carrier screening program was initially implemented in specific ethnogeographic groups with high risk of genetic disease [81,82]. A well-known example is the carrier screening program that was firstly introduced in the Jewish community in the 1970s with an attempt to reduce cases of Tay-Sachs disease [83]. Another typical example is the carrier screening program for β-thalassaemia, initially launched in Mediterranean countries and the Middle East where the disease is more prevalent and screening has become mandatory in Iran and Saudi Arabia [84]. These screening programs were highly successful and reduced incidence of Tay-Sachs disease and β-thalassemia in the respective groups by 80-90% [83,85].In contrast, The ACOG has recommend the genetic testing of CF, spinal muscular atrophy and sickle cell disease for across populations. Additional population-specific carrier screening programs include testing for carriers of Bloom syndrome, Gaucher disease Type I, and Canavan disease in Ashkenazi Jews [86].

As most population-based carrier screening programs are implemented in high-risk communities, this population-tailored strategy may fail to identify the risk of AR disorders at other ethnical groups due to differences in disease allele composition. For instance, up to 12% of newborn cases of sickle cell diseases are ethnically from the low-risk populations in California [87]. For example, *CFTR*, the gene whose disruption is causative for CF, is highly polymorphic with more than 2,000 variants that have been described. Importantly however, frequencies of the majority of these vary drastically among worldwide populations [31]. For instance, the ΔF508 mutation accounts for 70% of CF causative variants in Caucasian CF patients [88], whereas it only explains <30% of cases in African populations [89]. As a consequence, optimal pathogenic variant panels may differ vastly between populations and therefore hamper the predictive performance of current screening programs. Additionally, a substantial fraction of the AR disorder cases fail to be detected by the currently establish screening programs, as evidenced by the fact that only 80% of CF cases can be accounted for by screening panel recommended by ACMG and ACOG [6].

1.3.2.2.2 Expanded carrier screening programs

Conventional genetic screening programs only focus on a small selection of specific genetic diseases. However, the vast majority of rare genetic diseases is not covered in these panels, resulting in false negative assessments and underestimated disease risk [90]. Importantly, the advent of NGS facilitates allows for the first time to systematically screen for pathogenic variants of multiple genetic diseases across genome. Expanded carrier screening has therefore emerged in which hundreds of rare disorders are probed in addition to those monogenic disorders already included in conventional screening programs. Such a strategy was found to be cost-effective compared to screening programs focused on individual candidate variants [81,90]. Until 2017, at least 16 providers offered expanded carrier screening that meets analytical requirements, testing between 41 and 1,792 conditions [91]. The included conditions cover not only monogenic disorders but also provide guidance for the consideration of cognitive disability status and clinical interventions [92]. However, the genetic pathogenicity of approximately 50% of rare diseases remains to be elucidated and can as such currently not be included into expanded screening programs [93].

2 AIMS

2.1 GENERAL AIM

The overarching aim of this PhD project is to leverage population-scale sequencing data to facilitate precision medicine, specifically by discovering how genetic variability can serve as biomarker for treatment stratification and guide genetic screening programs.

2.2 STUDY-SPECIFIC AIMS

The specific aim of each constituent paper was to use publicly available sequencing data to:

Paper I:

Unravel the genetic landscape of the human *ABC* transporter superfamily and its ethnogeographic variability.

Paper II:

Explore the predictive value of *ABC* transporter variability for breast cancer chemotherapy.

Paper III:

Provide quantitative estimates of incidence and genetic complexity of 480 human AR disorders.

3 MAIN METHODS

3.1 NGS DATA OF POPULATION SEQUENCING PROJECTS

In **Paper I-III,** 138,632 individuals' NGS data of worldwide seven ethnic groups, whose ancestry constitutes Africans, Latinos, Ashkenazi Jewish, East Asians, South Asians, Finnish, non-Fin Europeans, were accessed from gnomAD [31]. These data are publicly available and there is neither publication nor use limitation.

3.2 NGS DATA OF CANCER GENOME SEQUENCING PROJECT

In **Paper II**, breast invasive carcinoma (BRCA), cell renal carcinoma (ccRCC) and liver hepatocellular carcinoma (HCC) were studied as they are among most common cancers reported. WGS data of peripheral and bulk tumor-adjacent healthy tissue from these cancer patients were accessed from the The Cancer Genome Atlas (TCGA) database for germline genetic variant analysis, with granted ethical permit from US National Institutes of Health. The corresponding clinical data were downloaded from the TCGA Pan-Cancer Clinical Data Resource [94]. The variant calling was performed with GATK v4.1 according to the best practice workflow [95,96].

3.3 VARIANT EFFECT PREDICTION

In **Paper I-III**, functionality of missense variants was predicted with a variety of best performed algorithms evaluated by Li et al [97]. In **Paper III** exclusively, variant pathogenicity was firstly annotated with ClinVar [98]. Additionally, the pathogenicity of the frameshift, splicing site, stop gained and start lost variants was predicted using LOFTEE plugin (https://github.com/konradjk/loftee) of Variant Effect Predictor [99].

3.4 DISEASE INCIDENCE ESTIMATION AND GENETIC COMPLEXITY

In **Paper III**, disease-gene associations were identified from the Online Mendelian Inheritance in Man (OMIM) database [75]. Genetic incidences of AR disorders were estimated as q^2 where q denotes the aggregated MAF of pathogenic variants of the causative gene(s) of each disorder. Additionally, 90% Jeffrey's confidence interval (CI) was calculated for putatively genetic incidence of each AR disorder using "ratesci" from R [100]. The genetic complexity of each studied disorder was evaluated by informedness index, calculated as $I = \max_v (D(v) - P(v))$ where v stands for an array of the MAF-sorted interrogated variants. The $D(v)$ denotes the fraction of the disease whose molecular pathogenesis is the interrogated variants included v, while $P(v)$ denotes the percentage of variant number of v to number of identified all pathogenic variants in a given disorder.

3.5 TERTIARY STRUCTURE ANALYSIS

Papers I and II systematically investigated the potential variant effect by mapping the variant on protein tertiary structures. The experimentally determined structures were accessed from Protein Data Bank (https://www.rcsb.org/) [101], while others were either from literature or modelled using Phyre2 [102]. PyMOL (v2.1.1) was employed for genetic variability visualization on the functional domains.

3.6 STATISTICS

In **Paper II**, best cutoff of variant burden among the cancer cohort were optimized with conditional inference tree using “partykit” package in R [103], while the potential dichotomization effect was examined with linear tail-restricted cubic spline modeling implemented in R “rms” package [104]. Associations between single variants and variant burden were identified with linear-by-linear association tests in R package “coin” [105]. The survival difference of different groups was computed using log-rank test. $P < 0.05$ was considered as statistically significant for all tests.

4 RESULTS AND DISCUSSION

4.1 PAPER I: ETHNOGEOGRAPHIC AND INTERINDIVIDUAL VARIABILITY OF HUMAN *ABC* TRANSPORTERS

Human ABC transporters mediate a spectrum of endogenous substrates such as peptides and sugars. Furthermore, ABC transporters translocate a spectrum of drug substrates, including HIV protease inhibitors and chemotherapy agents. Genetic polymorphisms in *ABC* genes can modulate the risk for inappropriate drug response or toxicity, and can underlie a variety of genetic diseases [106–109]. While the importance of *ABC* variants has been well recognized, previous studies focused almost exclusively on a few candidate polymorphisms and information on the overall profile of *ABC* variability had been lacking.

4.1.1 Interindividual differences in *ABC* transporter variability and their functional implications

Paper I analyzed the landscape and profile of genetic variability in the human *ABC* transporter superfamily. Across all 48 *ABC* transporters encoded in the human genome, 37,467 genetic variants were identified that impacted on protein primary structure, which included missense (n=33,340), splice site (n=924), indels (n=435), frameshifts (n=1,549), as well as variants resulting in the loss of the canonical start codon or the gain of a premature stop codon (n=1,219).

Of these variants, 19,309 variants were estimated to be deleterious and 17,524 missense variants were predicted as neutral by 5 state-of-the-art algorithms (Fig. 7A). Highest deleterious variant burden was in *ABCA13* (n=1,183), *ABCA7* (n=953) and *ABCA4* (n=865), while the least deleterious variants were discovered in *ABCB7* (n=43) and *ABCE1* (n=60; Fig. 7B). Importantly, the vast majority of putatively deleterious variants were found to be rare with MAF < 1% (19,266 out of 19,309; 99.7%). On average, individuals harbored most variants with functional consequences in *ABCB5* (n=1.8) and *ABCB1* (n=1.1; Fig. 7C). Of these, rs2032582, a variant reportedly associated with altered drug response and toxicity in different contexts of cancer chemotherapy [55,110,111], accounted for 80% of the functional genetic variability of *ABCB1*. By contrast, 24 of 48 ABC transporters harbored no common variants with putative functional relevance and the entire functional genetic variability was attributed to rare variations. Taken together, each individual was estimated to harbor on average 9.8 deleterious *ABC* variations of which 21% were owing to rare genetic variants (Fig. 7D). These findings were similar to other highly polymorphic gene classes, such as the *SLC* [53] and *SLCO* [54] transporter families, suggesting potentially important roles of rare genetic variants for *ABC* transporter function.

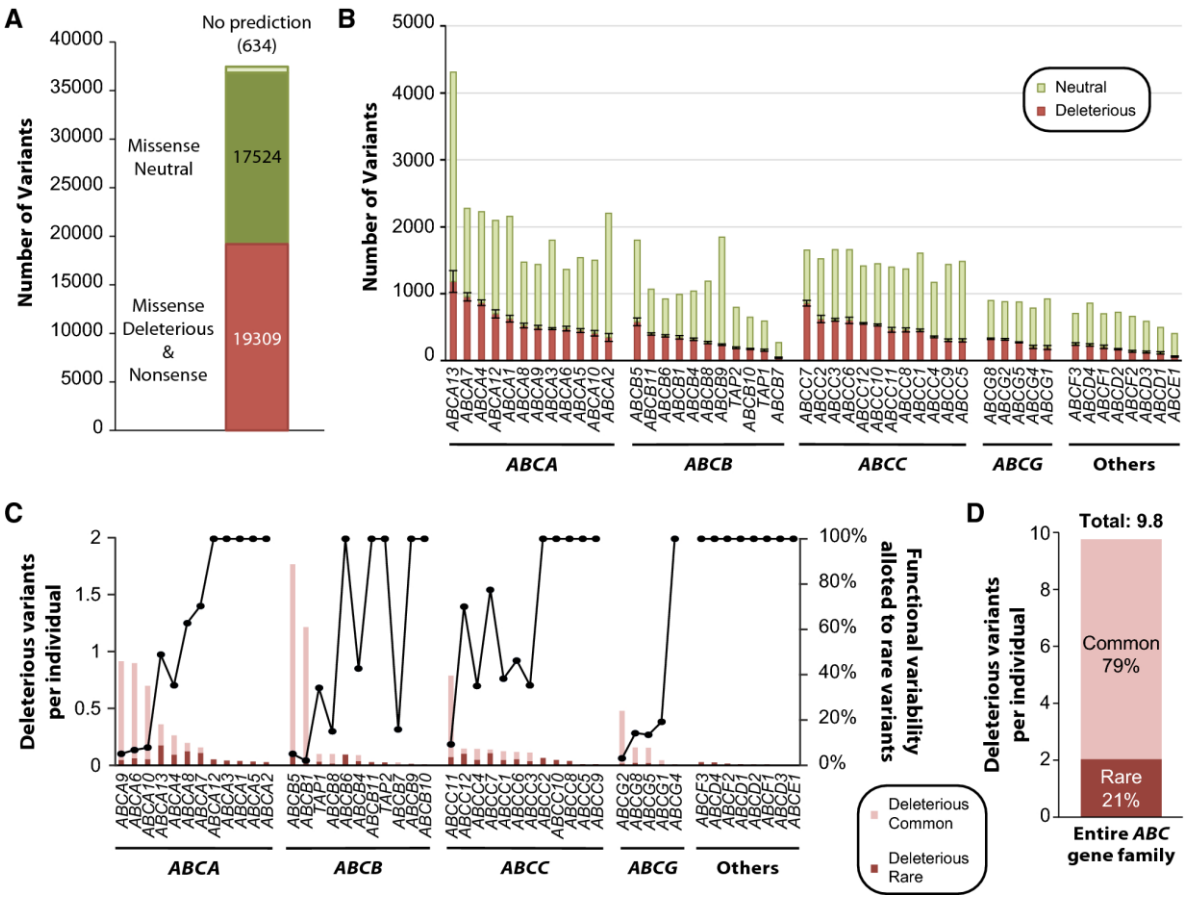

Fig 7. ABC functional genetic variability. ***A***, *Overview of variant functionality prediction.* ***B***, *Results of functionality prediction of each ABC gene.* ***C***, *Number of individual harbored ABC deleterious variants and corresponding contribution by rare genetic variability.* ***D***, *Estimated deleterious variant burden of ABC variants per individual.*

4.1.2 Ethnogeographic variability within the *ABC* transporter superfamily

Notably, 76% of the functional genetic variability was attributed to variants that were exclusively found in a single population, of which *ABCA7* and *ABCE1* harbored the lowest (70%) and highest (92%) fraction of population-specific variants, respectively (Fig. 8A-C). Vast majority of these population-specific variants were found in Europeans (n=6,815) and South Asians (n=2,413) whereas Finns (n=368) and Jews (n=136) were the least population-specific, likely at least in part due to the disproportionally large number of Europeans in the available data set. In contrast, only 0.3% of all putatively deleterious variants circulated among all populations studied.

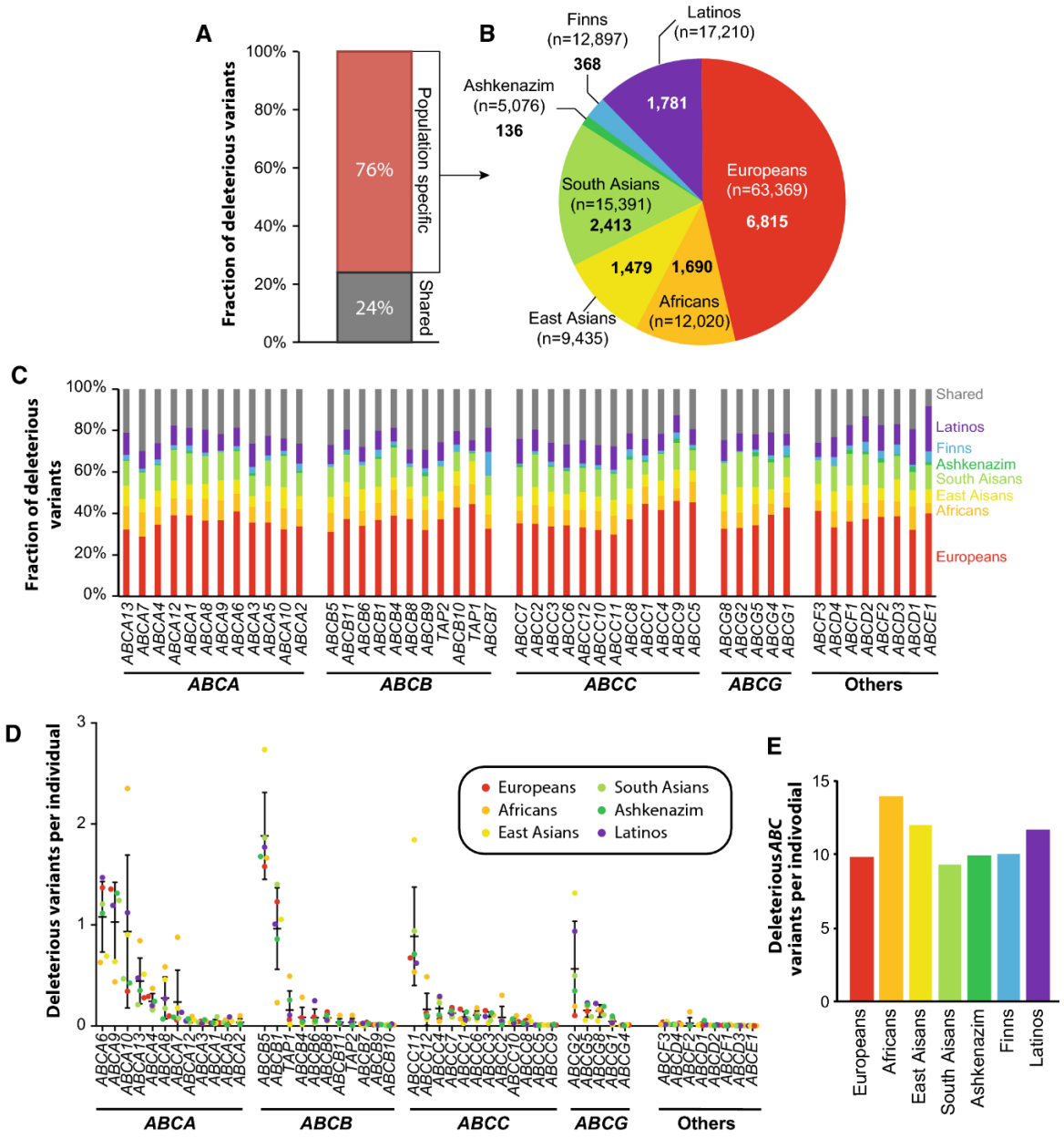

Fig 8. Population-specificity of putatively deleterious variants. ***A****, Percentage of population-specific and shared variants.* ***B****, Distribution of population-specific deleterious variants in each component population.* ***C****, Population-specificity of deleterious variants of each ABC gene.* ***D****, Estimated deleterious variant burden of ABC variants per individual of each population.* ***E****, Estimated overall ABC variant burden per individual of each population.*

Largest difference of each population was observed in *ABCA10* with each African supposedly harboring 2.4 putatively deleterious variants contrast to 0.3 variants in Europeans (Fig. 8D). Other significant interethnic differences were found in *ABCC11* (1.8 putatively deleterious variants observed in East Asians contrast to 0.5 in Africans), *ABCB1* (1.4 variants observed in South Asians contrast to 0.2 in Africans) and *ABCG2* (1.3 observed in East Asians contrast to 0.1 in Europeans). On the contrary, *ABCE1*, *ABCF*s and *ABCD*s harboured the least putatively deleterious variants and,

consequently, the lowest differences across populations. Across the entire gene superfamily, Africans harboured most putatively deleterious *ABC* variants (n=13.9 deleterious variants per individual) while the lowest numbers were predicted in South Asians (n=9.3; Fig. 8E). These results together revealed the extensive population-specificity of *ABC* genetic functional variability, and incentivized the use of population-specific strategies to profile *ABC* variability.

4.1.3 Conclusions

Paper I investigated the worldwide genetic variability of the *ABC* transporter superfamily using NGS data of >130,000 worldwide unrelated individuals. These analyses revealed that individuals harboured on average 9.3 to 13.9 putatively deleterious ABC transporter variants across the different populations, 21% of which were allotted to rare genetic variants and 76% of which were population-specific. As such, the study unveiled the drastic inter-individual and populational differences of ABC transporters on an unprecedented scale.

4.2 PAPER II: GERMLINE VARIANT BURDEN IN MULTIDRUG RESISTANCE TRANSPORTERS IS A THERAPY-SPECIFIC PREDICTOR OF SURVIVAL IN BREAST CANCER PATIENTS

Drug resistance constitutes the main clinical issue in oncology, accounting for approximately 90% anticancer treatment failures [44]. ABC transporter-mediated drug efflux is one important mechanism of chemotherapy resistance. *ABCB1*, *ABCC1* and *ABCG2* encode the most relevant multi-drug resistance transporters MDR1, MRP1 and BCRP, respectively. A limited number of *ABC* germline variants have been found related with altered drug response or toxicity across a range of cancers [110–112]. However, most studies used small heterogenous cohorts and, likely as a consequence, the identified associations commonly failed to replicate and results between studies were often conflicting. Furthermore, due to the focus on single variant associations, no study addressed the predictive value of the rare genetic variability in these genes.

4.2.1 Variant burden of *ABC* transporters predicts cancer prognosis

Germline variant and outcome data from the TCGA database were leveraged to evaluate whether *ABC* transporter variability can predict the prognosis of cancer patients. Common, previously reported variations were firstly examined that whether they could explain therapeutic outcomes using disease-specific survival (DSS) as a proxy. Notably however, none of the analysed common variations was associated with DSS. It was then hypothesized that drug resistance might be due to combinatorial effects of multiple variations with individually small effect sizes and thus shifted from a variant centric to a mutational burden model. To this end, permutation tests were employed to evaluate the power of variant burden cut-offs, while log-rank tests were employed for comparing DSS in high and low variant burden groups.

High variant burden of *ABCC1* correlated with significantly reduced DSS in BRCA patients (HR=3.22; 95%CI=[1.62-6.4]; log-rank p=0.00043, permutation p=0.031; Fig. 9). Furthermore, moderate associations were identified for *ABCB1* variability and decreased survival in ccRCC patients (HR=1.83, 95%CI=[1.0-3.37], log-rank p=0.047, permutation p=0.13), as well as for variant burden in *ABCC1* and increased DSS in HCC patients (HR=0.58, 95%CI=[0.35-0.98], log-rank p=0.039, permutation p=0.17).

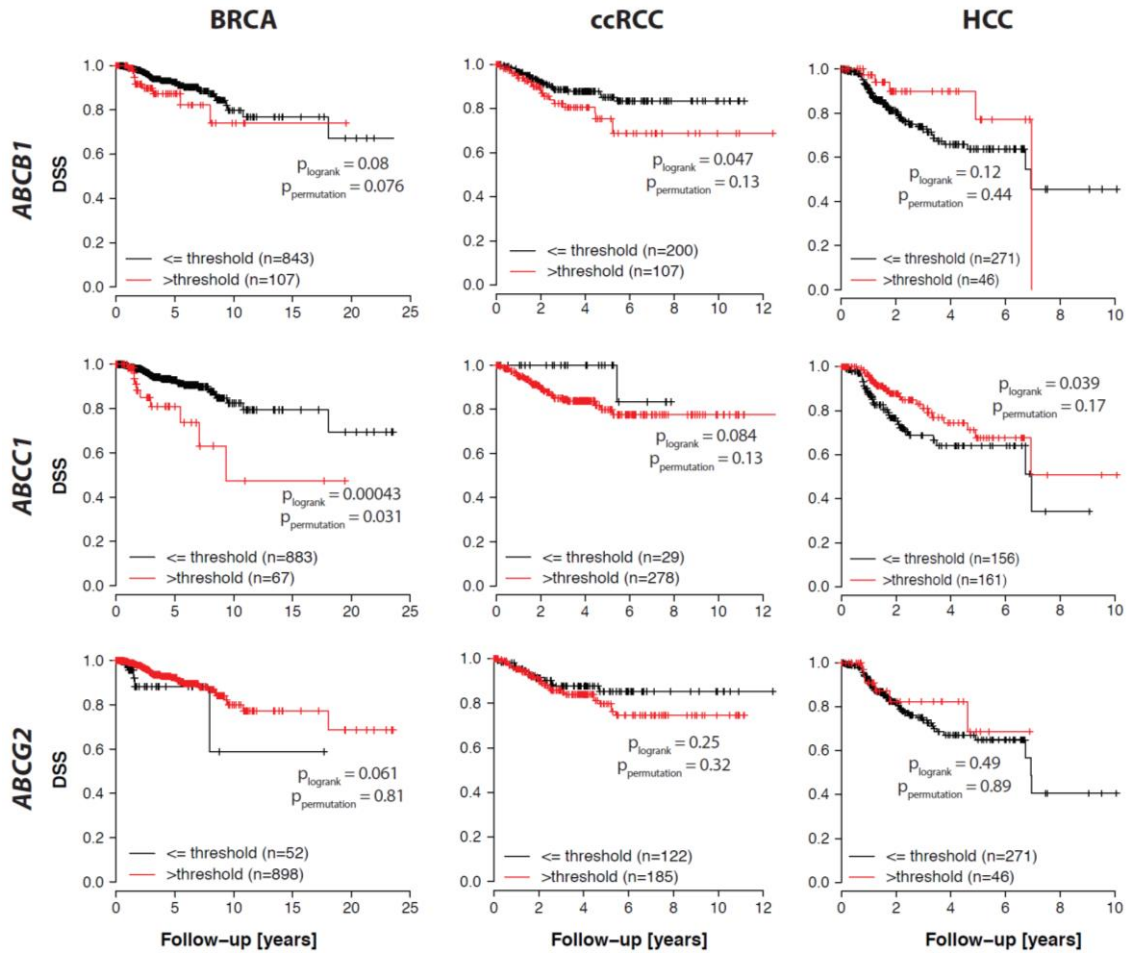

Fig 9. The impact of variant burden of multidrug resistance related ABC transporters on cancer DSS. *High variant burden of ABCC1 was found to strongly correlate reduced DSS in BRCA.*

4.2.2 Predictive value of *ABC* gene variant burden is drug-specific

Importantly, when these analyses were stratified by therapeutic regimen, it demonstrated strong drug-specific associations, providing additional mechanistic support (Table 1). Specifically, variant burden of *ABCC1* strongly predicted survival in the subgroup of BRCA patients treated with cyclophosphamide (HR=9.22, 95%CI=[1.83-46.36], log-rank p=0.0011) and doxorubicin (HR=4.57, 95%CI=[1.31-15.93], log-rank p=0.0088), two cytotoxic MRP1 substrates [113], whereas no significant association was found for tamoxifen-treated patients (log-rank p=0.13), a selective estrogen receptor modulator that inhibits the expression of estrogen response genes in the breast and is not known to be significantly transported by MRP1.

Table 1: Effect of variant burden of ABCB1, ABCC1 and ABCG2 on cancer DSS based on TCGA cohort.

	ABCB1		***ABCC1***		***ABCG2***	
	HR* [95% CI]	**Logrank test**	**HR* [95% CI]**	**Logrank test**	**HR* [95% CI]**	**Logrank test**
BRCA (n=960)	1.83 [0.92-3.64]	0.080	3.22 [1.62-6.40]	**0.00043**	0.43 [0.17-1.07]	0.061
Cyclophosphamide subgroup (n=238) ‡	1.99 [0.41-9.61]	0.38	9.22 [1.83-46.36]	**0.0011**	0.30 [0.04-2.41]	0.23
Doxorubicin subgroup (n=329) ‡	2.36 [0.78-7.18]	0.12	4.57 [1.31-15.93]	**0.0088**	0.25 [0.07-0.86]	**0.018**
Tamoxifen subgroup (n=238) ‡	0.34 [0.02-6.4]	0.23	3.07 [0.67-14.06]	0.13	0.70 [0.04-13.41]	0.56
ccRCC (n=314)	1.83 [1.0-3.37]	**0.047**	4.85 [0.67-35.2]	0.084	1.46 [0.76-2.82]	0.25
HCC (n=325)	0.49 [0.2-1.23]	0.12	0.58 [0.35-0.98]	**0.039**	0.75 [0.32-1.74]	0.49

***** The references of the HRs were defined as low variant groups.

‡ Variant burden thresholds were computed individually in the groups treated with different drugs.

4.2.3 *ABC* variants potentially associated with drug-specific resistance localize to functionally important transporter domains

MRP1 (encoded by *ABCC1*) consists of two transmembrane domains (TMDs) containing one nucleotide-binding domain (NBD) each (Fig. 10A). Of the 34 *ABCC1* variants potentially associated with drug-specific resistance, multiple encoded amino acid locates in functionally important domains. The variants encode G231C and R230Q are situated inside the Lasso motif connecting domains between TMD0 and TMD1 [114], specifically in a section functionally related with membrane attachment and transport activity [115]. R433S is located in a region close to the P-pocket responsible for coordination of the glutathione moiety while R1248 is positioned adjacent to the H-pocket containing lipid tail [114]. C1047S and R1066W are situated inside the cytoplasmic loop CL6 and its interaction with the 6-amino group of ATP has been shown to be essential for the transmission of conformational changes by ATP binding to the TMDs [116].

Similarly, rare variants associated with reduced survival were also found in functionally important regions of BCRP (encoded by *ABCG2*) (Fig. 10B). Contrary to MRP1, BCRP is a homodimer composed of two ABCG2 encoded half-transporters. T542A and C438R are positioned close to the translocation pore and might interfere with substrate translocation [117]. Q141K is located in and increase the instability of NBD, leading to significant decrease of protein expression of BCRP [118].

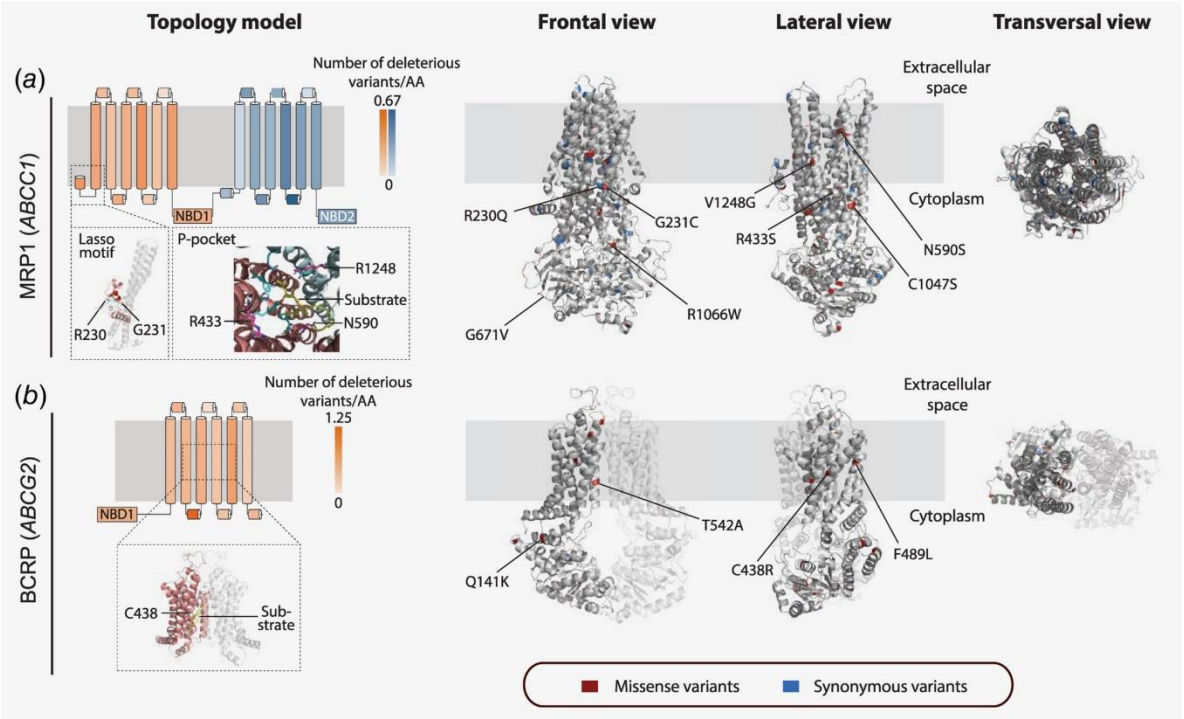

Fig 10. Structural mapping of variants enriched in or associated with high risk group of the BRCA cohort. ***A****, MRP1 consists of 2 NBDs.* ***B****, BCRP is homodimer of ABCG2-encoded protein. The color of the topology models was shaded according to the enrichment of deleterious variants based on variant effect prediction. The function domains were amplified with 3D structure with the located variants. All variants were additionally marked in integral tertiary structures.*

4.2.4 Conclusions

Paper II studied the predictive value of genetic variability in chemoresistance related transporters *ABCB1*, *ABCC1*, *ABCG2* on cancer chemotherapy outcome. Importantly, the germline variant burden of *ABCC1* constituted a drug-specific biomarker of DSS in BRCA patients whereas no significant associations were found for individual variants. These findings provided a new perspective and indicated the added value of variant burden testing using NGS compared to conventional SNP interrogations.

4.3 PAPER III: THE PREVALENCE, GENETIC COMPLEXITY AND POPULATION-SPECIFIC FOUNDER EFFECTS OF HUMAN AUTOSOMAL RECESSIVE DISORDERS

4.3.1 Comprehensive identification of variants associated with human autosomal recessive diseases

Overall, 508 genes causative for 450 AR diseases were analysed (Fig. 11). Only monogenic diseases with well-established AR inheritance pattern were included. Across all genes a total of 574,524 were identified, of which variations in low confidence regions and variants with known benign effects were excluded. 5,188 with established pathogenicity were taken forward for future analyses, whereas the remaining 553,018 variants were analysed for the putative variant effects using an array of computational tools.

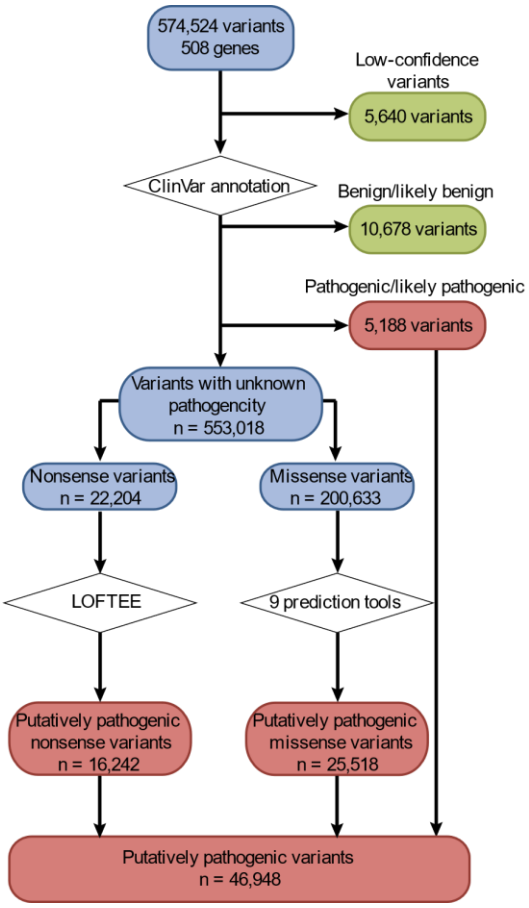

Fig 11. Workflow of variant pathogenicity prediction.

Functionality of nonsense variants (n=22,204) was predicted using LOFTEE, whereas the pathogenicity of missense variants (n=200,633) were evaluated using 9 partly orthogonal algorithms (see methods) which demonstrated the consistently best performance on missense pathogenicity prediction in independent datasets [97], and only those variants with unanimous predictions were considered as pathogenic. Combined, this workflow resulted in the identification of overall 46,948 putatively pathogenic variants.

4.3.2 Validation of the pathogenicity prediction model

To validate the approach for pathogenic variant selection and effect prediction, the estimated disease prevalence of each disease was compared with reported epidemiology data. Importantly, the model aligned overall well with established disease incidences (Pearson correlation r=0.68, $p < 0.0001$; Fig. 12). For instance, carrier rates in population-scale sequencing data allowed to accurately predict disease prevalence of sickle cell anemia in Africans (1 in 474 predicted *vs* 1 in 500 reported), Gaucher disease (1 in 1,019 predicted *vs* 1 in 750 reported), Canavan disease (1 in 8,564 predicted *vs* 1 in 9,950 reported) and Tay-Sachs disease (1 in 3,864 predicted

vs 1 in 3,500 reported) in Ashkenazi Jews, as well as A1AT deficiency (1 in 1,891 predicted *vs* 1 in 2,500 reported) in Europeans. The model also fitted well for CF (1 in 1,666 predicted *vs* 1 in 3,000 reported) in Europeans and Stargardt disease (1 in 2,333 predicted *vs* 1 in 9,000 reported) in general populations.

Notable outliers included type II citrullinemia, which was overestimated with reported incidence 1 in 165,000 and the estimated is 1 in 1,815, as well as type I tyrosinemia (1 in 196,167 contrasted to 1 in 16,000 reported) and AFP deficiency (1 in 2,948,001 contrast to 1 in 105,000 reported). Based on these data it was concluded that the *in silico* prediction that includes the computational prediction of rare, otherwise uncharacterized variants, overall accurately reflected the genetic basis for most AR disorders and thus constitutes a useful tool for the estimation of population-specific incidences for AR Mendelian diseases for which this information is not available.

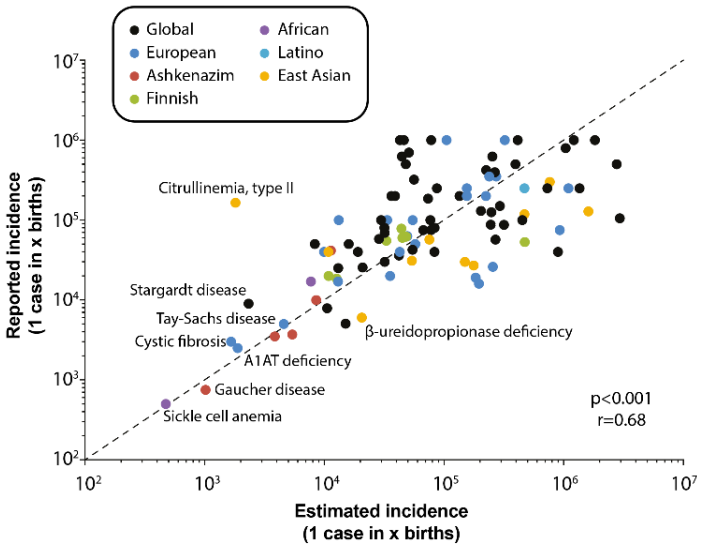

Fig 12. Validation of pathogenicity prediction with epidemiology data.

4.3.3 Genetic complexity of autosomal recessive disorders

Next, the genetic complexity of the 450 investigated AR disorders was analysed. The informedness (see methods), defined as the maximal difference between the percentage of disease explained by genetic factors and the fraction of the number of variants used to explain pathogenicity, was employed to estimate the genetic complexity. This metric can evaluate the additive performance of given putatively pathogenic variants compared to a uniform variant allocation. As such, a higher informedness value indicates a higher predictive power of a selected variant panel for molecular disease diagnosis.

Interestingly, informedness as a proxy for genetic complexity differed drastically between diseases, ranging from 0 to 0.96 (Fig. 13A). Congenital short bowel syndrome was a representative disease for high genetic complexity, as evidenced by a low informedness value of 0.25, indicating that the pathogenic variants underlying

the disease were similarly prevalent (Fig. 13B). In contrast, CPS1 deficiency (I=0.48) was moderately complex, with 70% of disease incidence being explained by 23% of the identified pathogenic variants. Similarly in CF (I=0.69), 10% of putatively pathogenic variants predicted 78% disease occurrence. The least genetically complex disease was hemochromatosis type I (I=0.96), in which a single variant only (rs1800562, C282Y) explained 99% of disease cases. Across all 450 diseases combined, 2.2% pathogenic variants (n=1,026) were sufficient to predict 50% AR disease risk (Fig. 13C).

The diseases were later analysed together with informedness, number of variants interrogated to achieve informedness, and the overall number of pathogenic variants for each disease. Next, it was speculated that informedness could provide a useful measure for genetic founder effects, as the diseases were disproportionally enriched in a few specific variations. Based on the analyses, 29 AR diseases were estimated to have strong founder effect, such as CF and Stargardt disease (Fig. 13D). These results align with previous reports [119,120].

The population-specific disease complexity was analysed next (Fig. 13E). The genetic complexity of AR diseases was overall lower in relatively isolated populations, such as Ashkenazi Jews and Finns, whereas complexity was markedly higher in Latinos, which are genetically more heterogenous. AR disorders were most genetically complex in Europeans, likely due to the larger sampling within Europe [31] that results in the capture of variants that might be missed in ancestries represented by smaller samples.

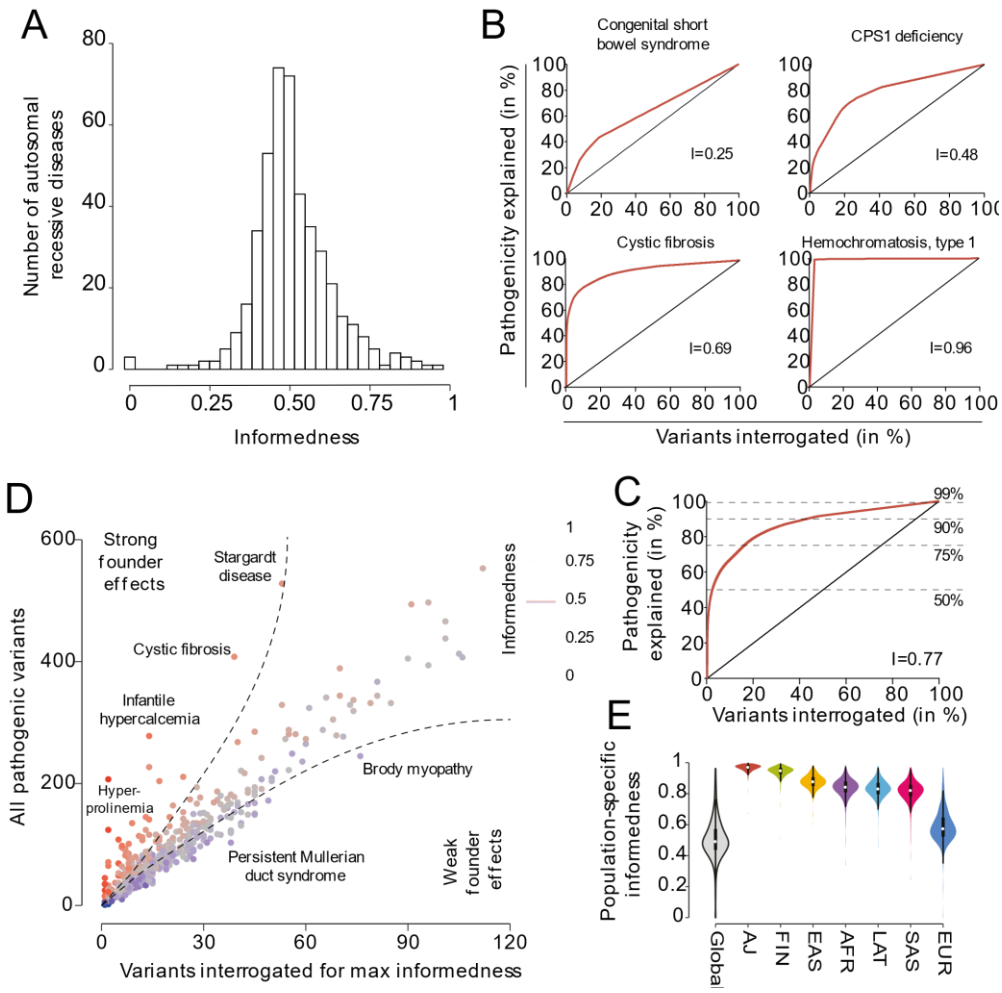

Fig 13. Genetic complexity of studied 450 AR disorders. ***A****, Overview of genetic complexity of studied AR disorders. The higher informedness indicates lower genetic complexity of a given disease.* ***B***, Selected examples of differently complex diseases. ***C****, Informedness calculated from all 450 disorders.* ***D***, *Disease complexity integrating number of pathogenic variants, informedness and informedness derived number of variants.* ***E****, Population-specific informedness of all 450 disorders.*

4.3.4 Conclusions

Paper III systematically identified the landscape of pathogenic variants across 508 genes associated with human AR disorders and used this extensive data set to estimate the genetic prevalence and complexity of 450 AR disorders. The findings demonstrated that population-scale sequencing data can provide a powerful resource for molecular disease genetics even in the absence of functional annotations. Furthermore, these results revealed that AR disorders differ drastically in their genetic constitution and population-specificity with important implications for the design of genetic screening programs.

5 THESIS CONCLUSIONS AND FUTURE PERSPECTIVES

5.1 CONCLUSIONS

In summary, this PhD thesis leveraged population-scale NGS data for multiple applications in precision genomic medicine. Specifically, the work aspires to provide cases that elucidate the translational potential or different use cases in order to incentivize the further exploitation of the available sequencing projects in the rapidly evolving space of genome-guided cancer chemotherapy and carrier screening.

The papers together highlighted the importance of genomics on precision medicine, exemplified its clinical associations in two contexts of medicine. It's envisioned that more studies based on analysis of NGS projects could contribute substantially to precision medicine in different subareas.

5.2 FUTURE PERSPECTIVES

5.2.1 Pharmacogenomics in cancer treatment

As described above, genetic factors account for 20-30% of interindividual differences in drug toxicity and response [2]. Importantly however, effects of common *ABC* transporter variants on chemotherapeutic outcomes appear to be highly context dependent and commonly fail to reproduce between studies, which renders them poor biomarkers. As such, the additional findings reported in this thesis indicate that the incorporation of comprehensive NGS data improves predictive models, thus incentivizing the transition from candidate SNP interrogations to NGS-guided predictions. These advances due to increased data availability is paralleled by developments in computational algorithms to predict the functionality of rare or novel variants without available experimental characterization data. Combined, in the future, comprehensive germline sequencing data analysed with advanced bioinformatic tools that allow to include the entire repertoire of genetic variation into their prediction models might emerge as a new strategy in precision oncology.

5.2.2 Carrier screening program

The currently established genetic carrier screening programs capture only a subset of disease-associated genes and even within those only a limited set of variations [86,121]. As such, substantial proportion of AR disease carriers remain unidentified. The rapid evolution of computational algorithms enables the progressively accurate identification of pathogenic variants of interest. Thus, it is envisioned that their integration with comprehensive sequencing-based approaches holds much promise

to substantially enhance the performance of screening panels in a cost-effective manner.

ACKNOWLEDGEMENT

I would primarily thank my main supervisor Assoc. Prof. Dr. **Volker Lauschke**. It's my fortune to join Dr. Lauschke's group as a PhD student. I have learnt solid knowledge regarding biomedicine, and result interpretation as well as the scientific writing under his supervision. I'm also extremely grateful to my co-supervisor Dr. **Isabel Barragan**, who supported me to participate in bioinformatics courses in the beginning of my PhD study.

I'm grateful for the support from **China Scholarship Council**.

I acknowledge Dr. **Stefan Winter**, Dr. **Florian Büttner**, Dr. **Elke Schaeffeler**, Prof. **Matthias Schwab** for their contributions, specifically the detailed survival analysis presented in Paper II.

I also appreciate the colleagues in Lauschke group for sharing knowledge, which include **Yitian Zhou,** Dr. **Shane Wright, Carolina Dagli Hernandez**, Dr. **Julianna Kele Olovsson**, **Aurino Kemas**, **Stefania Koutsilieri**, **Joanne Shen**, **Despoina-Christina Sismanoglou**, **Nuria Vilarnau**, Dr. **Sonia Youhanna** and Dr. **Reza Zandi Shafagh**.

I extend my thanks to Prof. **Magnus Ingelman-Sundberg** and colleagues in his group for joint seminars and journal clubs during my PhD education, which includes Dr. **Marcela Franco**, Dr. **Riina Harjumäki**, **Inger Johansson**, Dr. **Marin Jukic**, **Vlasia Kastrinou-Lampou**, Dr. **Katharina Klöditz**, **Mikael Kozyra**, **Souren Mkrtchian**, **Åsa Nordling**, Dr. **Christopher Pridgeon** and Dr. **Sara Redensek**.

I would also acknowledge the program of **The Swedish National Graduate School in Medical Bioinformatics**, where I have acquired substantial knowledge of bioinformatics, especially from Dr. **Lars Arvestad**, Dr. **Lukas Käll**, Dr. **Arne Elofsson**, Dr. **Samuel Flores**, as well as my bioinformatics mentor in the program, Dr. **Carsten Daub**.

I would also thank other colleagues in Department of Physiology and Pharmacology. Many thanks to head Prof. **Håkan Westerblad** and PhD study director Assoc. Prof. **Kent Jardemark** for their in-time help during my PhD study. And I also thank Prof. **Elisabet Stener-Victorin** and Prof. **Carl Johan Sundberg**, and Dr. **Qiaolin Deng** for their useful suggestions during my PhD study. Many thanks to the departmental PhD study administrator **Sofia Pettersson** for always being available for the practical procedures.

Last but not least, I always thank to my parents for their selfless love and support. Without these I will never have the courage to start the PhD education.

REFERENCES

1. Ashley EA. Towards precision medicine. Nature Reviews Genetics. 2016;17:507–22.

2. Lauschke VM, Ingelman-Sundberg M. Prediction of drug response and adverse drug reactions: From twin studies to Next Generation Sequencing. European Journal of Pharmaceutical Sciences. 2019;130:65–77.

3. Matthaei J, Tzvetkov M, Strube J, Sehrt D, Sachse-Seeboth C, Hjelmborg J, et al. Heritability of Caffeine Metabolism: Environmental Effects Masking Genetic Effects on CYP1A2 Activity but Not on NAT2: Heritability of caffeine metabolism. Clinical Pharmacology & Therapeutics. 2016;100:606–16.

4. Matthaei J, Brockmöller J, Tzvetkov M, Sehrt D, Sachse-Seeboth C, Hjelmborg J, et al. Heritability of metoprolol and torsemide pharmacokinetics. Clinical Pharmacology & Therapeutics. 2015;98:611–21.

5. Bauwens M, Garanto A, Sangermano R, Naessens S, Weisschuh N, De Zaeytijd J, et al. ABCA4-associated disease as a model for missing heritability in autosomal recessive disorders: novel noncoding splice, cis-regulatory, structural, and recurrent hypomorphic variants. Genetics in Medicine. 2019;21:1761–71.

6. Ioannou L, McClaren BJ, Massie J, Lewis S, Metcalfe SA, Forrest L, et al. Population-based carrier screening for cystic fibrosis: a systematic review of 23 years of research. Genetics in Medicine. 2014;16:207–16.

7. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics. 2010;11:446–50.

8. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019;177:26–31.

9. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538:161–4.

10. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics. 2013;45:580–5.

11. Puangpetch A, Vanwong N, Nuntamool N, Hongkaew Y, Chamnanphon M, Sukasem C. CYP2D6 polymorphisms and their influence on risperidone treatment. Pharmacogenomics and Personalized Medicine. 2016;9:131–47.

12. Zhou Y, Ingelman-Sundberg M, Lauschke V. Worldwide Distribution of Cytochrome P450 Alleles: A Meta-analysis of Population-scale Sequencing Projects. Clinical Pharmacology & Therapeutics. 2017;102:688–700.

13. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016;107:1–8.

14. Somatic Mutations in Cerebral Cortical Malformations. The New England Journal of Medicine. 2014;2.

15. National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC), United Kingdom Inflammatory Bowel Disease Genetics Consortium, International Inflammatory Bowel Disease Genetics Consortium, Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of GWAS loci identifies independent

rare variants associated with inflammatory bowel disease. Nature Genetics. 2011;43:1066–73.

16. Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnology. 2008;26:1135–45.

17. Schuster SC. Next-generation sequencing transforms today's biology. Nature Methods. 2008;5:16–8.

18. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.

19. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016;17:333–51.

20. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biology. 2020;21:30.

21. Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Research. 2018;46:2159–68.

22. Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. Long reads: their purpose and place. Human Molecular Genetics. 2018;27:R234–41.

23. Bleidorn C. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. Systematics and Biodiversity. 2016;14:1–8.

24. Mantere T, Kersten S, Hoischen A. Long-Read Sequencing Emerging in Medical Genetics. Frontiers in Genetics. 2019;10:426.

25. Höijer I, Tsai Y-C, Clark TA, Kotturi P, Dahl N, Stattin E-L, et al. Detailed analysis of *HTT* repeat elements in human blood using targeted amplification-free long-read sequencing. Human Mutation. 2018;39:1262–72.

26. Ammar R, Paton TA, Torti D, Shlien A, Bader GD. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. F1000Res. F1000Research; 2015;4:17–17.

27. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431:931–45.

28. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.

29. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.

30. Exome Aggregation Consortium, Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.

31. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.

32. Collins RL, Brand H, Karczewski KJ, Zhao X, et al. A structural variation reference for medical and population genetics. Nature. 2020;581:444–51.

33. Minikel EV, Karczewski KJ, Martin HC, Cummings BB, et al. Evaluating drug targets through human loss-of-function genetic variation. Nature. 2020;581:459–64.

34. Ehmann F, Caneva L, Prasad K, Paulmichl M, Maliepaard M, Llerena A, et al. Pharmacogenomic information in drug labels: European Medicines Agency perspective. The Pharmacogenomics Journal. 2015;15:201–10.

35. Varnai R, Szabo I, Tarlos G, Szentpeteri LJ, Sik A, Balogh S, et al. Pharmacogenomic biomarker information differences between drug labels in the United States and Hungary: implementation from medical practitioner view. The Pharmacogenomics Journal. 2020;20:380–7.

36. Relling MV, Evans WE. Pharmacogenomics in the clinic. Nature. 2015;526:343–50.

37. The SEARCH Collaborative Group. *SLCO1B1* Variants and Statin-Induced Myopathy — A Genomewide Study. New England Journal of Medicine. 2008;359:789–99.

38. Anderson HD, Crooks KR, Kao DP, Aquilante CL. The landscape of pharmacogenetic testing in a US managed care population. Genetics in Medicine. 2020;22:1247–53.

39. Amstutz U, Farese S, Aebi S, Largiadèr CR. Dihydropyrimidine dehydrogenase gene variation and severe 5-fluorouracil toxicity: a haplotype assessment. Pharmacogenomics. 2009;10:931–44.

40. Offer SM, Fossum CC, Wegner NJ, Stuflesser AJ, Butterfield GL, Diasio RB. Comparative Functional Analysis of DPYD Variants of Potential Clinical Relevance to Dihydropyrimidine Dehydrogenase Activity. Cancer Research. 2014;74:2545–54.

41. van Kuilenburg ABP, Meijer J, Mul ANPM, Meinsma R, Schmid V, Dobritzsch D, et al. Intragenic deletions and a deep intronic mutation affecting pre-mRNA splicing in the dihydropyrimidine dehydrogenase gene as novel mechanisms causing 5-fluorouracil toxicity. Human Genetics. 2010;128:529–38.

42. Relling MV, Schwab M, Whirl-Carrillo M, Suarez-Kurtz G, Pui C, Stein CM, et al. Clinical Pharmacogenetics Implementation Consortium Guideline for Thiopurine Dosing Based on *TPMT* and *NUDT 15* Genotypes: 2018 Update. Clinical Pharmacology & Therapeutics. 2019;105:1095–105.

43. Takano M, Sugiyama T. UGT1A1 polymorphisms in cancer: impact on irinotecan treatment. Pharmacogenomics and Personalized Medicine. 2017;10:61–8.

44. Longley D, Johnston P. Molecular mechanisms of drug resistance. The Journal of Pathology. 2005;205:275–92.

45. Holohan C, Van Schaeybroeck S, Longley DB, Johnston PG. Cancer drug resistance: an evolving paradigm. Nature Reviews Cancer. 2013;13:714–26.

46. Vasiliou V, Vasiliou K, Nebert DW. Human ATP-binding cassette (ABC) transporter family. Human Genomics. 2008;3:281.

47. Fletcher JI, Haber M, Henderson MJ, Norris MD. ABC transporters in cancer: more than just drug efflux pumps. Nature Reviews Cancer. 2010;10:147–56.

48. Piehler AP, Hellum M, Wenzel JJ, Kaminski E, Haug K, Kierulf P, et al. The human ABC transporter pseudogene family: Evidence for transcription and gene-pseudogene interference. BMC Genomics. 2008;9:165.

49. Ambudkar SV, Kimchi-Sarfaty C, Sauna ZE, Gottesman MM. P-glycoprotein: from genomics to mechanism. Oncogene. 2003;22:7468–85.

50. Kimchi-Sarfaty C, Oh JM, Kim I-W, Sauna ZE, Calcagno AM, Ambudkar SV, et al. A “Silent” Polymorphism in the MDR1 Gene Changes Substrate Specificity. Science. 2007;315:525–8.

51. Kozyra M, Ingelman-Sundberg M, Lauschke VM. Rare genetic variants in cellular transporters, metabolic enzymes and nuclear receptors can be important determinants of interindividual differences in drug response. Genetics in Medicine. 2017;19:20–9.

52. Fujikura K, Ingelman-Sundberg M, Lauschke VM. Genetic variation in the human cytochrome P450 supergene family: Pharmacogenetics and Genomics. 2015;25:584–94.

53. Schaller L, Lauschke VM. The genetic landscape of the human solute carrier (SLC) transporter superfamily. Human Genetics. 2019;138:1359–77.

54. Zhang B, Lauschke VM. Genetic variability and population diversity of the human SLCO (OATP) transporter family. Pharmacological Research. 2019;139:550–9.

55. Dulucq S, Bouchet S, Turcq B, Lippert E, Etienne G, Reiffers J, et al. Multidrug resistance gene (MDR1) polymorphisms are associated with major molecular responses to standard-dose imatinib in chronic myeloid leukemia. Blood. 2008;112:2024–7.

56. Wojnowski L, Kulle B, Schirmer M, Schlüter G, Schmidt A, Rosenberger A, et al. NAD(P)H Oxidase and Multidrug Resistance Protein Genetic Polymorphisms Are Associated With Doxorubicin-Induced Cardiotoxicity. Circulation. 2005;112:3754–62.

57. Sparreboom A. Diflomotecan pharmacokinetics in relation to ABCG2 421C>A genotype*1. Clinical Pharmacology & Therapeutics. 2004;76:38–44.

58. Deeken JF, Cormier T, Price DK, Sissung TM, Steinberg SM, Tran K, et al. A pharmacogenetic study of docetaxel and thalidomide in patients with castration-resistant prostate cancer using the DMET genotyping platform. The Pharmacogenomics Journal. 2010;10:191–9.

59. Lee C. Vive la difference! Nature Genetics. 2005;37:660–1.

60. Varela MA, Amos W. Heterogeneous distribution of SNPs in the human genome: Microsatellites as predictors of nucleotide diversity and divergence. Genomics. 2010;95:151–9.

61. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nature Reviews Genetics. 2006;7:85–97.

62. Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. Human Molecular Genetics. 2010;19:R131–6.

63. Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, et al. Natural genetic variation caused by small insertions and deletions in the human genome. Genome Research. 2011;21:830–9.

64. Collins F, Drumm M, Cole J, Lockwood W, Vande Woude G, Iannuzzi M. Construction of a general human chromosome jumping library, with application to cystic fibrosis. Science. 1987;235:1046–9.

65. Warren S, Zhang F, Licameli G, Peters J. The fragile X site in somatic cell hybrids: an approach for molecular cloning of fragile sites. Science. 1987;237:420–3.

66. Bondeson M-L, Dahl N, Malmgren H, Kleijer WJ, Tönnesen T, Carlberg B-M, et al. Inversion of the IDS gene resulting from recombination with IDS-related sequences in a common cause of the Hunter syndrome. Human Molecular Genetics. 1995;4:615–21.

67. Osborne LR, Li M, Pober B, Chitayat D, Bodurtha J, Mandel A, et al. A 1.5 million–base pair inversion polymorphism in families with Williams-Beuren syndrome. Nature Genetics. 2001;29:321–5.

68. Kurotaki N, Harada N, Shimokawa O, Miyake N, Kawame H, Uetake K, et al. Fifty microdeletions among 112 cases of Sotos syndrome: Low copy repeats possibly mediate the common deletion. Human Mutation. 2003;22:378–87.

69. Park J-H, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nature Genetics. 2010;42:570–5.

70. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.

71. Forgetta V, Manousaki D, Istomine R, Ross S, Tessier M-C, Marchand L, et al. Rare Genetic Variants of Large Effect Influence Risk of Type 1 Diabetes. Diabetes. 2020;69:784–95.

72. Del-Aguila JL, Koboldt DC, Black K, Chasse R, Norton J, Wilson RK, et al. Alzheimer's disease: rare variants with large effect sizes. Current Opinion in Genetics & Development. 2015;33:49–55.

73. Soutar AK. Rare genetic causes of autosomal dominant or recessive hypercholesterolaemia. IUBMB Life. 2010;62:125–31.

74. Diaz-Horta O, Duman D, Foster J, Sırmacı A, Gonzalez M, Mahdieh N, et al. Whole-Exome Sequencing Efficiently Detects Rare Mutations in Autosomal Recessive Nonsyndromic Hearing Loss. Janecke AR, editor. PLoS ONE. 2012;7:e50628.

75. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Research. 2015;43:D789–98.

76. on behalf of the European Society of Human Genetics (ESHG), Henneman L, Borry P, Chokoshvili D, Cornel MC, van El CG, et al. Responsible implementation of expanded carrier screening. European Journal of Human Genetics. 2016;24:e1–12.

77. Srinivasan BS, Evans EA, Flannick J, Patterson AS, Chang CC, Pham T, et al. A universal carrier test for the long tail of Mendelian disease. Reproductive BioMedicine Online. 2010;21:537–51.

78. Mouzan MIE, Salloum AAA, Herbish ASA, Qurachi MM, Omar AAA. Consanguinity and major genetic disorders in Saudi children: a community-based cross-sectional study. Annals of Saudi Medicine. 2008;5.

79. Abouelhoda M, Sobahy T, El-Kalioby M, Patel N, Shamseldin H, Monies D, et al. Clinical genomics can facilitate countrywide estimation of autosomal recessive disease burden. Genetics in Medicine. 2016;18:1244–9.

80. Vallance H, Ford J. Carrier Testing for Autosomal- Recessive Disorders. Critical Reviews in Clinical Laboratory Sciences. 2003;40:473–97.

81. van der Hout S, Holtkamp KC, Henneman L, de Wert G, Dondorp WJ. Advantages of expanded universal carrier screening: what is at stake? European Journal of Human Genetics. 2017;25:17–21.

82. Zlotogora J. Population programs for the detection of couples at risk for severe monogenic genetic diseases. Human Genetics. 2009;126:247–53.

83. Kaback M, Lim-Steele J, Dabholkar D, Brown D, Levy N, Zeiger K. Tay-Sachs disease--carrier screening, prenatal diagnosis, and the molecular era. An international perspective, 1970 to 1993. The International TSD Data Collection Network. JAMA. 1993;270:2307-2315.

84. Cousens NE, Gaff CL, Metcalfe SA, Delatycki MB. Carrier screening for Beta-thalassaemia: a review of international practice. European Journal of Human Genetics. 2010;18:1077–83.

85. Cao A. Results of programmes for antenatal detection of thalassemia in reducing the incidence of the disorder. Blood Reviews. 1987;1:169–76.

86. Gross SJ, Pletcher BA, Monaghan KG. Carrier screening in individuals of Ashkenazi Jewish descent. Genetics in Medicine. 2008;10:54–6.

87. Shafer FE, Lorey F, Cunningham GC, Klumpp C, Vichinsky E, Lubin B. Newborn Screening for Sickle Cell Disease: 4 Years of Experience from California's Newborn Screening Program. Journal of Pediatric Hematology/Oncology. 1996;18:36–41.

88. Beaudet AL. Genetic Testing for Cystic Fibrosis. Pediatric Clinics of North America. 1992;39:213–28.

89. Stewart C, Pepper MS. Cystic Fibrosis in the African Diaspora. Annals of the American Thoracic Society. 2017;14:1–7.

90. Nazareth SB, Lazarin GA, Goldberg JD. Changing trends in carrier screening for genetic disease in the United States: Expanded carrier screening. Prenatal Diagnosis. 2015;35:931–5.

91. Chokoshvili D, Vears D, Borry P. Expanded carrier screening for monogenic disorders: where are we now?: Expanded carrier screening for monogenic disorders. Prenatal Diagnosis. 2018;38:59–66.

92. Edwards JG, Feldman G, Goldberg J, Gregg AR, Norton ME, Rose NC, et al. Expanded Carrier Screening in Reproductive Medicine—Points to Consider: A Joint Statement of the American College of Medical Genetics and Genomics, American College of Obstetricians and Gynecologists, National Society of Genetic Counselors, Perinatal Quality Foundation, and Society for Maternal-Fetal Medicine. Obstetrics & Gynecology. 2015;125:653–62.

93. Boycott KM, Rath A, Chong JX, Hartley T, Alkuraya FS, Baynam G, et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. The American Journal of Human Genetics. 2017;100:695–705.

94. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018;173:400-416.e11.

95. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011;43:491–8.

96. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018;201178.

97. Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, et al. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Research. 2018;46:7793–804.

98. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research. 2014;42:D980–5.

99. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016;17:122.

100. Laud PJ. Equal-tailed confidence intervals for comparison of rates: Equal-tailed confidence intervals for comparison of rates. Pharmaceutical Statistics. 2017;16:334–48.

101. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research. 2000;28:235–42.

102. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols. 2015;10:845–58.

103. Hothorn T, Zeileis A. Partykit: a modular toolkit for recursive Partytioning in R. Journal of Machine Learning Research. 2015;16:3905–9.

104. Helmreich JE. Regression Modeling Strategies with Applications to Linear Models, Logistic and Ordinal Regression and Survival Analysis (2nd Edition). Journal of Statistical Software. 2016;70.

105. Hothorn T, Hornik K, van de Wiel MA, Zeileis A. A Lego System for Conditional Inference. The American Statistician. 2006;60:257–63.

106. DeStefano GM, Kurban M, Anyane-Yeboa K, Dall'Armi C, Di Paolo G, Feenstra H, et al. Mutations in the Cholesterol Transporter Gene ABCA5 Are Associated with Excessive Hair Overgrowth. Schmidt-Ullrich R, editor. PLoS Genetics. 2014;10:e1004333.

107. Lopes-Pacheco M. CFTR Modulators: Shedding Light on Precision Medicine for Cystic Fibrosis. Frontiers in Pharmacology. 2016;7:275.

108. Roberts RL, Wallace MC, Phipps-Green AJ, Topless R, Drake JM, Tan P, et al. ABCG2 loss-of-function polymorphism predicts poor response to allopurinol in patients with gout. The Pharmacogenomics Journal. 2017;17:201–3.

109. Vulsteke C, Lambrechts D, Dieudonné A, Hatse S, Brouwers B, van Brussel T, et al. Genetic variability in the multidrug resistance associated protein-1 (ABCC1/MRP1) predicts hematological toxicity in breast cancer patients receiving (neo-)adjuvant chemotherapy with 5-fluorouracil, epirubicin and cyclophosphamide (FEC). Annals of Oncology. 2013;24:1513–25.

110. Chang H, Rha SY, Jeung H-C, Im C-K, Ahn JB, Kwon WS, et al. Association of the ABCB1 gene polymorphisms 2677G>T/A and 3435C>T with clinical outcomes of paclitaxel monotherapy in metastatic breast cancer patients. Annals of Oncology. 2009;20:272–7.

111. Megías-Vericat JE, Montesinos P, Herrero MJ, Moscardó F, Bosó V, Rojas L, et al. Impact of *ABC* single nucleotide polymorphisms upon the efficacy and toxicity of induction chemotherapy in acute myeloid leukemia. Leukemia & Lymphoma. 2017;58:1197–206.

112. Chaturvedi P, Tulsyan S, Agarwal G, Lal P, Agarwal S, Mittal RD, et al. Influence of ABCB1 genetic variants in breast cancer treatment outcomes. Cancer Epidemiology. 2013;37:754–61.

113. Deeley RG, Cole SPC. Substrate recognition and transport by multidrug resistance protein 1 (ABCC1). FEBS Letters. 2006;580:1103–11.

114. Johnson ZL, Chen J. Structural Basis of Substrate Recognition by the Multidrug Resistance Protein MRP1. Cell. 2017;168:1075-1085.e9.

115. Bakos E, Evers R, Calenda G, Tusnady GE, Szakacs G, Varadi A, et al. Characterization of the amino-terminal regions in the human multidrug resistance protein (MRP1). Journal of Cell Science. 2000;113:4451.

116. Johnson ZL, Chen J. ATP Binding Enables Substrate Release from Multidrug Resistance Protein 1. Cell. 2018;172:81-89.e10.

117. Manolaridis I, Jackson SM, Taylor NMI, Kowal J, Stahlberg H, Locher KP. Cryo-EM structures of a human ABCG2 mutant trapped in ATP-bound and substrate-bound states. Nature. 2018;563:426–30.

118. Woodward OM, Tukaye DN, Cui J, Greenwell P, Constantoulakis LM, Parker BS, et al. Gout-causing Q141K mutation in ABCG2 leads to instability of the nucleotide-binding domain and can be corrected with small molecules. Proceedings of the National Academy of Sciences. 2013;110:5223–8.

119. Wood Klinger K. Cystic fibrosis in the Ohio Amish: Gene frequency and founder effect. Human Genetics. 1983;65:94–8.

120. Chacón-Camacho OF, Granillo-Alvarez M, Ayala-Ramírez R, Zenteno JC. ABCA4 mutational spectrum in Mexican patients with Stargardt disease: Identification of 12 novel mutations and evidence of a founder effect for the common p.A1773V mutation. Experimental Eye Research. 2013;109:77–82.

121. Watson MS, Cutting GR, Desnick RJ, Driscoll DA, Klinger K, Mennuti M, et al. Cystic fibrosis population carrier screening: 2004 revision of American College of Medical Genetics mutation panel. Genetics in Medicine. 2004;6:387–91.